计及负荷不确定性的强化学习实时定价策略 *

王菁祺，高　岩†，吴志强，李仁杰

(上海理工大学 管理学院，上海 200093)

摘　要：面对当前电力系统的负荷不确定、新能源并网与“双碳”目标等现状，在充分考虑供需双方福利前提下，建立了智能电网背景下考虑负荷不确定与碳交易的实时定价模型。并基于强化学习能够处理变量复杂性、非凸非线性问题优点，采用强化学习中 Q 学习算法对模型进行迭代求解。首先，将用户与供电商实时交互过程转换为强化学习框架对应的马尔可夫决策过程。其次，通过智能体在动态环境中的反复探索表示用户与供电商的信息交互。最后，通过强化学习中的 Q 学习算法寻找最优值即最大社会福利值。仿真结果表明，所提实时定价策略能够有效提升社会福利，降低碳排放总量，这验证了所提模型和算法的有效性。

关键词：实时定价；强化学习；马尔可夫决策过程；负荷不确定；“双碳”目标

中图分类号：TP391.9　　doi: 10.19734/j.issn.1001-3695.2022.02.0069

Real-time pricing strategy based on reinforcement learning with load uncertainty

Wang Jingqi, Gao Yan†, Wu Zhiqiang, Li Renjie

(*Business School*, *University of Shanghai for Science & Technology*, *Shanghai* 200093, *China*)

Abstract: Facing the current situation of load uncertainty, new energy grid integration, and “dual carbon” target in the power system, the paper established a real-time pricing model considering load uncertainty and carbon trading in the context of the smart grid with full consideration of the welfare of both supply side and user side. Based on the advantages that reinforcement learning can handle variable complexity, non-convex, and nonlinear problems, this paper used the Q-learning algorithm in reinforcement learning to solve the model iteratively. Firstly, this paper transformed the real-time interaction process between the user and the power supplier into a Markov decision process corresponding to the reinforcement learning framework. Secondly, the process represented the information interaction between the user and the power supplier as the iterative exploration of the agent in a dynamic environment. Finally, this paper found the optimal value by the Q-learning algorithm in reinforcement learning, i. e. , the maximal social welfare value. The simulation results show that the proposed real-time pricing strategy can effectively enhance social welfare and reduce total carbon emissions, which verifies the feasibility and effectiveness of the proposed model and algorithm.

Key words: real-time pricing; reinforcement learning; Markov decision-making process; load uncertainty; “dual carbon” goal

0　引言

在智能电网系统中，电力和信息的双向流动能够兼顾电力系统经济、高效、环境友好等目标。随着新能源发电商的深入普及，给发电系统带来了更大的不确定性。围绕着发电商、分布式新能源、碳交易市场与用户需求，需求侧管理将带来大量的产业机会。

随着信息通信与智能终端的发展，电力市场中电价的波动加剧，将增加普通用户参与电力系统调节的意愿。对电力系统需求侧进行管理能够有效对电力消耗削峰填谷，优化用电方式，提高电力系统的稳定性与安全性。需求响应(demand response, DR)是需求侧管理的解决方案之一。现有需求响应策略[1~3]通常分为激励型需求响应(incentive-based DR, IBDR)和价格型需求响应(price-based DR, PBDR)。价格型需求响应通过电价的调整使得用户改变其用电模式；激励型需求响应则向用户提供固定或随时间变化的激励费用。通过考虑用户的行为,许多研究使用基于价格的需求响应，而实时定价是价格型需求响应的重要研究方向，该策略通过直接控制电力价格以调整用户侧负荷需求，旨在通过提供实时电价有效地平抑用户的用电需求。

文献[4]首次提出了以社会福利最大化为目标的实时定价模型，模型同时考虑到供电商利润和用户福利，采用分布式梯度下降法求解，数值仿真验证了模型可实现削峰填谷，同时对用户和供电商两方均有益。在此基础上，以社会福利最大化作为目标函数的实时定价模型被广泛应用。文献[5]采用了光滑化方法对现有实时定价中常用的二次分段效用函数进行光滑化处理，并仿真得到用户效用。文献[6]以极小化峰谷差为目标建立实时定价优化模型，并提出一种依赖在线电量波动的同步扰动随机逼近算法。文献[7]将区块链引入实时定价模型，能够有效地提高微网可再生能源的利用率。同时用户也作为独立节点参与到电网决策中，应用区块链交易可充分提高用户用电的精准性和社会总福利。文献[8]将社会福利最大化模型与微电网进行有效结合，建立了一个计及不确定性的双层优化模型，并使用 PSO-BBA 算法进行求解，并通过与确定性函数的对比，能够更好地起到削峰填谷的作用。文献[9]在社会福利最大化模型上对最小供电量约束的作用进行了讨论，引入有效成本函数并提出了对偶在线算法，实现了模型的改进。文献[10]将实时定价问题表述为非合作博弈问题，并利用分布式在线算法进行求解，对用户交互过程进行了更加精准的描述。

收稿日期：2022-02-28；**修回日期：**2022-04-19　**基金项目：**国家自然科学基金资助项目(72071130)

作者简介：王菁祺(1997-)，男，河南平顶山人，博士研究生，主要研究方向为智能电网实时定价、机器学习；高岩(1962-)，男(通信作者)，黑龙江五常人，教授，博导，博士，主要研究方向为智能电网实时定价等(gaoyan@usst.edu.cn)；吴志强(1997-)，男，安徽合肥人，硕士研究生，主要研究方向为系统工程、决策分析；李仁杰(1992-)，男，江苏泰州人，博士研究生，主要研究方向为智能电网实时定价、机器学习.

表 1　符号说明

Tab. 1　Symbol description

符号	符号描述
$\mathcal{N}^R / \mathcal{N}^L / \mathcal{N}$	某区域居民/大型/总用户集合
T	所有时段的集合
A / S	强化学习动作空间/状态空间集合
$D_{t,n}^{basic} / D_{t,n}^{flex}$	用户 n 在 t 时段基本/可削减负荷需求
$X_{t,n}^{basic} / X_{t,n}^{flex}$	用户 n 在 t 时段基本/可削减负荷
π_0	基准电价
$c_n^{\min} / c_n^{\max}$	不同用户的最小/最大电力价格系数
$X_{t,n}$	用户 n 在 t 时段的总负荷
$\tilde{X}_{t,n}$	用户 n 在 t 时段实际总负荷
$\delta_{t,n}$	用户 n 在 t 时段总负荷的随机变量
$\sigma_{t,n}$	$\delta_{t,n}$ 的方差
$p_{t,n}$	用户 n 在 t 时段的电价
$\epsilon_{t,n}$	用户价格弹性系数
α_n / β_n	用户效用参数
L_t^e	传统能源供电商 t 时段供电量
$a_t / b_t / c_t$	电力成本系数
L_t^r	风光新能源供电商 t 时段供电量
P_t^{PV}	光伏发电 t 时段实际输出量
P_t^{WT}	风力发电 t 时段实际输出量
P_{PV}^{rated}	光伏发电额定输出功率
G_C / G_{PV}	光伏工作点实际/标准辐射强度
η_{PV}	光伏功率温度系数
$T_C / T_{PV,T}$	光伏发电实际/参考温度
N_{PV}	光伏发电设备数量
P_{WT}^{rated}	风力发电机额定输出功率
v / v_{rated}	风力发电实际风速/额定风速
v_{in} / v_{out}	风力发电切入/切出风速
N_{WT}	风力发电设备数量
$L_t^{e,min} / L_t^{e,\max}$	传统能源供电商在 t 时段的最小/最大供电量
$\delta_{RE} / \sigma_{RE}$	新能源设备维护损失成本系数
δ_e / δ_r	传统能源/新能源单位发电碳配额分配率
p_e	单位碳排放权的价格
$\alpha_e / \beta_e / \lambda_e$	传统能源发电的单位电量碳排放系数
∂ / γ	强化学习学习率/折现因子

从优化方法来看，上述实时定价策略大致分为基于梯度优化算法的[4~7]与基于元启发式优化算法的[8~11]两类。前者如共轭梯度法、牛顿法等，具有计算效率高的特点，但如果模型中存在非线性、非光滑函数或者机会约束等难以处理的情况，具有较好全局搜索能力的元启发式算法如遗传算法、粒子群算法等，大部分与给定的模型高度独立，可以很好地解决前者的问题。另一方面，现有定价策略往往预先确定模型的各项参数且集中式算法较多，在某种程度上没有考虑到负荷不确定性情况且对于隐私安全缺乏相应的保护措施。面对大规模批量数据时会出现运算速度过慢、可靠性较低等问题，创新实时定价机制具有重要的理论意义和现实意义。

从时间关联性上来看，上述研究主要将实时电价问题分为多个单时段问题予以考虑[4~9]，每个时段没有充分考虑整体的状态转移特性而独立存在，对于实时电价模型交互过程描述的精确性有待提高，忽略了用户用电和供电商供电的前后关联性，而马尔可夫决策过程可以使用状态转移矩阵描述负荷前后阶段的关系，可以充分考虑时段的关联性。文献[11,13]基于马尔可夫过程研究实时定价问题，考虑了参数已知与未知两种情况，并验证了模型的合理性与算法的可行性。

上述实时定价研究大多依赖于分析模型和确定性规则的传统算法。近年来，强化学习取得了新的进展。与传统优化算法不同，强化学习可在动态环境中探索一些随机行动并从经验中学习，从而可为求解复杂系统决策提供重要支持。强化学习简洁明了且使用奖励函数来评估决策行为，通过强化学习可得到问题有效的解决策略且结果具有收敛性。强化学习应用于许多领域，例如游戏控制，计算机视觉等[12]。而对于电力系统的强化学习研究具有较为广阔的前景，在电力系统需求侧管理中采用强化学习将有效扩展新的负荷侧用电模式[13]。

近年来，强化学习算法在需求侧管理中的应用主要有两类，第一类是站在消费者立场，面对供电商的定价策略设计有效的响应模式以最大化消费者的利益[15]。第二种是站在公用事业公司的立场通过设计有效的策略提高社会福利，从而有效提高包含用户侧与供电侧在内的福利[14,16]。Lu 等[14]首次将强化学习方法应用于需求侧管理，提出了分级电力市场的实时定价算法，将供电商与用户的交互表示为马尔可夫决策过程，从而动态确定最优电价。文献[15]使用强化学习获取需求响应中特定设备的能量调度，并在调度期间最大化用户的回报。文献[16]应用强化学习框架与需求响应策略，考虑到工业用户与供电商的交互过程，实现供电商长期收益最大化。文献[17]应用强化学习方法并将微电网视为一个智能体，微电网之间可通过单独选择能源交易策略，目标是最大化各个微电网的平均收益。文献[18]提出了一种基于神经网络和强化学习算法的多微电网能源管理方法，运营商通过深度神经网络来预测各微网的功率交换，通过蒙特卡洛方法求解得到零售定价策略，使得运营商达到利润最大化与需求侧的峰均比最小化目标，提高用电可靠性。

然而上述基于强化学习的需求侧管理研究缺乏对社会福利、碳交易与负荷不确定情况的整体考虑[13~17]。基于上述分析，有必要对实时定价模型进行相应扩展，使用强化学习算法求解实时定价模型有显著优势，考虑到供电商产电所带来的碳排放权以及碳排放交易所带来的成本或收益，本文通过引入碳排放权交易促进新能源消纳，进而助力“双碳”目标的实现。

本文主要工作如下：

a)考虑到含传统能源供电商与新能源供电商组成的供电商系统以及居民用户和大型用户组成的用户系统，并充分表示考虑了供需双方的福利，目标为社会福利最大。

b)通过引入强化学习框架将用户与供电商之间的交互过程表述为马尔可夫决策过程，利用智能体与环境即供电商与全体用户的迭代过程学习和获取最优的实时定价策略。

c)将实时定价模型与强化学习的各要素进行了对应，并充分考虑了负荷不确定等情况，从而实现了对模型更加精细地刻画。

d)通过引入碳交易，有效地提高电力系统新能源的消纳率，对推动能源可持续绿色发展有重要的现实意义。

1　系统模型

考虑一种包含两类供电商和若干个不同类型终端用户的智能电网系统(系统框架如图 1 所示，符号说明部分见表 1)，其中供电商包含传统能源供电商与新能源供电商，新能源供电由风力发电与光伏发电构成，同时由于新能源供电本身的间歇性、不稳定性等特性，供电商无法控制其每时段出力值，需根据风光机组特性及当日天气作出当日各时段的预测。即用户用电由新能源供电优先供应，从而促进新能源的消纳。用户侧考虑居民用户与大型用户，居民用户能源消耗为日常生活用电，而工商业等大型用户的能源消耗往往是为了更高的利润。

假设用户和供电商直接通过智能电表进行双向信息交互，

即供电商可以通过智能电表获取用户的电力消耗情况，同时用户可根据智能电表获取下一时段供电商提供的价格信号。即供电商侧通过实时定价策略实现利润最大化，用户侧通过需求响应策略动态调整他们的能源需求从而降低购电成本，因此可以根据用户侧的负荷需求和供电侧产电成本交互动态调整电价。

$\mathcal{N}^R=\{1,2,3,\cdots,m\}$ 表示居民用户集合，$\mathcal{N}^L=\{m+1,m+2,\cdots,n\}$ 表示大型用户集合。$\mathcal{N}$ 代表全体用户集合，$\mathcal{N}=\mathcal{N}^R\cup\mathcal{N}^L$。供电商与用户电力交互以一天为周期，将其分为 t 一个时段，$t\in T$，$T=\{1,2,3,\cdots,t\}$ 是所有时段的集合，模型假设 $t=24$，即价格每小时更新一次。同时本文考虑到负荷不确定与碳排放权交易情景，建立了社会福利最大化目标下的实时定价模型。

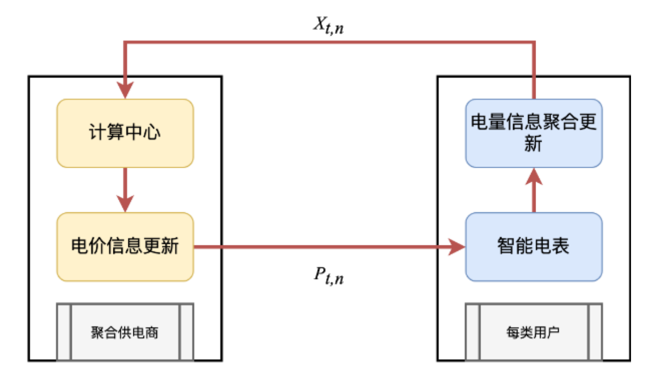


图 1　系统框架

Fig. 1　System framework

1.1　用户侧模型

一般情况下，用户在电力市场所需要的电量和对相同电量的消耗后的效用值不尽相同。根据用户侧负荷优先级和需求特征，本文假设用户负荷配置分为两类，基本负荷与可削减负荷[19]。在特定时段内固定需求的负荷称为基本负荷，可以灵活调配使用时间的负荷则称为可削减负荷。用户可通过灵活调节空调、热水器等可削减负荷实现需求响应。在需求响应中，供电商通过价格的动态调整引导用户改变该时段的用电需求，从而实现供需平衡。

1.1.1 负荷函数

假设用户基本负荷需要严格满足，即不能通过需求响应调控该类负荷，例如生活必需用电。

用户 n 在 t 时段的基本负荷 $X_{t,n}^{basic}$ 与基本负荷需求量 $D_{t,n}^{basic}$ 关系如下：

$$X_{t,n}^{basic}=D_{t,n}^{basic}. \tag{1}$$

同时，考虑可灵活调配使用时间及功率的负荷，称为可削减负荷。可削减负荷与当前时间的电价以及当前用户的价格弹性系数有关，用户 n 在 t 时段可削减负荷的定义为[16]：

$$X_{t,n}^{flex}=D_{t,n}^{flex}(1-\epsilon_{t,n}\frac{p_{t,n}-c_n^{\min}\pi_0}{c_n^{\min}\pi_0}),\forall n\in\mathcal{N},\forall t\in T \tag{2}$$

$$c_n^{\min}\pi_0\leqslant p_{t,n}\leqslant c_n^{\max}\pi_0,\forall n\in\mathcal{N} \tag{3}$$

其中 $p_{t,n}$ 表示用户 n 在 t 时段需支付的电价，$D_{t,n}^{flex}$ 表示用户 n 在 t 时段可削减负荷需求量，$\epsilon_{t,n}>0$ 为用户 n 在 t 时段的价格弹性系数。价格的升高导致用户实际负荷小于预期需求量，同时供电商的电价也应该在一个固定的区间内，π_0 为基准电价，$c_n^{\min}$ 和 $c_n^{\max}$ 分别代表电力价格系数的下界与上界，不同类型用户的电力价格系数也不同。通过电力价格约束可保证供需双方以合理的价格进行电力交易[20]。

令 $X_{t,n}$ 表示用户 n 在 t 时段中的电力总负荷，包含基本负荷与可削减负荷，表示如下：

$$X_{t,n}=X_{t,n}^{basic}+X_{t,n}^{flex},\forall n\in\mathcal{N},\forall t\in T \tag{4}$$

同时由于现实环境的变化，考虑到用户侧负荷的随机性，电力装置通常会面临负荷波动。在考虑负荷波动情况下，用户 n 在 t 时段总负荷 $\tilde{X}_{t,n}$ 为

$$\tilde{X}_{t,n}=X_{t,n}+\delta_{t,n},\forall n\in\mathcal{N},\forall t\in T \tag{5}$$

其中 $\delta_{t,n}\sim N(0,\sigma_{t,n}{}^2)$ 是一个随机变量且服从正态分布，$\sigma_{t,n}>0$，刻画了用户侧负荷的不确定性[21]。

由于可削减负荷具有对电价敏感的特性，合理调度可削减负荷能够有效实现电网削峰填谷。

1.1.2 效用函数

微观经济学中，效用函数 $U(x)$ 可以刻画用户的满意程度。假设每一个用户对于不同电价的行为均是独立的，对负荷的需求有着不同的偏好。弹性系数 β_n 可以有效体现不同用户的需求偏好，根据实际情况，效用函数 $U(x)$ 需要满足：a) $\frac{\partial U(x)}{\partial x}>0$；b) $U(0,\beta)=0$ $(\forall\beta>0)$。现有的实时定价模型中，用户的效用函数常用二次函数表示[22]，即用户 $n\in\mathcal{N}$ 在 $t\in T$ 时段的效用函数 $U(\tilde{X}_{t,n})$ 可以表示：

$$U(\tilde{X}_{t,n})=\begin{cases}\beta_n\tilde{X}_{t,n}-\frac{\alpha_n}{2}(\tilde{X}_{t,n})^2, & 0\leqslant\tilde{X}_{t,n}\leqslant\frac{\beta_n}{\alpha_n}\\ \frac{(\beta_n)^2}{2\alpha_n}, & \tilde{X}_{t,n}>\frac{\beta_n}{\alpha_n}\end{cases} \tag{6}$$

其中 $\tilde{X}_{t,n}$ 为用户 n 在 t 时段总负荷。$\beta_n>0$，$\alpha_n>0$ 为用户效用参数[23,24]，在实际应用中应根据历史数据和用户调研来估计。不同类型用户效用的变化程度可通过参数 α_n，β_n 刻画。

与居民用户类似，在一定负荷消费范围内，大型用户效用会随着电力消费水平的增加而增加，当达到预先定义的最大负荷量时，效用将保持恒定，但用户侧负荷通常不会达到饱和状态。

综上所述，用户侧福利可以表示为用户在当前时段效用值函数减去所支付成本的期望。令 π^C 表示用户侧福利，则用户侧福利函数表示如下:

$$\pi^C=E[\sum_{t\in T}\sum_{n\in\mathcal{N}}(U(\tilde{X}_{t,n})-p_{t,n}\tilde{X}_{t,n})] \tag{7}$$

1.2　供电商模型

供电商按照用户的电力需求向用户提供电力，从而实现电力的生产与传输。近年来，风电、光伏等新能源的接入大大增加了电力系统的随机性。令 L_t^e 和 L_t^r 分别代表传统能源与新能源供电商在 t 时段的发电量，由于供电商总供电量需要覆盖所有用户的需求，L_t^e 需满足机组发电区间约束，则 L_t^e 和 L_t^r 需满足如下约束：

$$\sum_{n\in\mathcal{N}}E(\tilde{X}_{t,n})\leq L_t^e+L_t^r \tag{8}$$

$$L_t^{e,\min}\leq L_t^e\leq L_t^{e,\max} \tag{9}$$

其中 $L_t^{e,\min}$ 和 $L_t^{e,\max}$ 分别代表传统能源供电商在 t 时段的最小与最大发电量。

1.2.1 传统能源供电商

假设传统能源供电商成本主要来源于化石能源消耗和运行维护，传统能源发电成本函数是一个单调增加的且严格凸的函数，目前普遍采用二次函数表示供电商发电成本[23]，供电商在 t 时段的发电成本函数 $C_t^i(L_t^e)$ 如下:

$$C_t^i(L_t^e)=a_t(L_t^e)^2+b_tL_t^e+c_t \tag{10}$$

其中 L_t^e 指传统能源供电商在 t 时段内提供的总电量，$a_t>0$，$b_t\geq0$，$c_t\geq0$ 为预设参数。

1.2.2 新能源供电商

由于光照强度、风速等自然资源的间歇性，新能源供电的输出功率存在较大的不确定性，若系统可调配容量不足，则将造成弃风弃光现象，大大破坏系统的稳定性。针对此情况，本文假设新型能源供电不具有存储功能且新能源供电与发电之间没有耦合约束，同时供电商优先使用新能源供电以提高新能源消纳率。

光伏发电输出主要取决于到达地面的太阳辐射强度、环境温度和光伏模块本身的特性。光伏发电机组在 t 时段内的实际输出功率[8]为

$$P_t^{PV} = P_{PV}^{\text{rated}}(G_c / G_{PV})(1-\eta_{PV,i}(T_c - T_{PV,T}))N_{PV} \tag{11}$$

其中 P_{PV}^{rated} 表示额定光伏输出功率； G_C 表示工作点的辐射强度； G_{PV} 表示标准辐射强度； η_{PV} 表示功率温度系数； T_C 表示工作点的电池温度； $T_{PV,T}$ 表示参考温度； N_{PV} 代表光伏发电设备数量。

风力发电输出功率与当前时段内实际风速有关。一般来说，风速波动服从瑞利分布，风力发电机组在 t 时段内实际输出功率为[8]：

$$P_t^{WT} = \begin{cases} 0, & v < v_{in}, v > v_{out} \\ \dfrac{v - v_{in}}{v_{rated} - v_{in}} P_{WT}^{rated} N_{WT}, & v_{in} \le v < v_{rated} \\ P_{WT}^{rated} N_{WT}, & v_{rated} \le v < v_{out} \end{cases} \tag{12}$$

其中 v 代表实际风速； v_{rated} 是额定风速； v_{in} 和 v_{out} 分别代表切入和切出风速； P_{WT}^{rated} 表示额定输出功率； N_{WT} 代表风力发电设备数量。

新能源供电包含风力发电输出与光伏发电输出， L_t^r 表示新能源供电在 t 时段的总输出功率，表述如下：

$$L_t^r = P_t^{PV} + P_t^{WT}, \forall t \in T \tag{13}$$

由于新能源发电成本可忽略不计，假设新能源供电商成本来自于后期运行维护的费用，本文使用二次成本函数表示 t 时段新能源设备运行过程中维护损失成本[25]，表述如下：

$$C_t^{RE}(L_t^r) = \delta_{RE}(L_t^r)^2 + \sigma_{RE} L_t^r, \forall t \in T \tag{14}$$

其中 $\delta_{RE} > 0$， $\sigma_{RE} \ge 0$ 为新能源设备维护损失成本系数。

1.2.3 碳交易模型

碳交易机制下通过碳排放权交易可促进电力系统“双碳”目标的实现，在碳排放权交易体系下，国家会根据供电商的发电总量分配相应的碳排放配额。若供电商的实际排放量小于分配的排放额度，则可将剩余额度在市场上出售获利；若供电商的实际碳排放量超过了分配的排放额度，须在市场上购买出超出部分的碳排放权，并由此产生碳过排放成本[26]。

供电商可通过传统能源发电与可再生能源发电获得碳排放权，发电机组在 t 时段分配的碳排放配额 E_t^D 如下：

$$E_t^D = \delta_e L_t^e + \delta_r L_t^r \tag{15}$$

其中 δ_e 和 δ_r 分别代表传统能源与新能源发电的单位碳排放配额分配率。

并考虑传统能源发电作为碳排放量来源，传统能源发电机组在 t 时段实际碳排放量如下所示[27]：

$$E_t^C = \alpha_e (L_t^e)^2 + \beta_e L_t^e + \lambda_e \tag{16}$$

其中 $\alpha_e, \beta_e, \lambda_e$ 为传统能源发电商产单位电量的碳排放系数。

综上，可得 t 时段碳交易成本 C_t^E 计算公式如下：

$$C_t^E = p_e(E_t^C - E_t^D) \tag{17}$$

其中 p_e 是市场上每单位碳排放权的交易价格， $C_t^E \ge 0$ 表示碳排放过量产生的碳交易成本，反之为碳交易收益。

考虑包含传统能源供电商以及新能源供电商构成的供电商集合，在不考虑供电商之间电力交互的情况下，供电商通过向用户出售电力获得售电收益，同时由于存在非清洁能源发电会带来相应碳排放量从而产生碳交易成本，社会偏好使用环境友好型的清洁能源，减少碳排放，促进电力系统的可持续发展。

定义供电商福利为售电收入与成本之差的期望，供电商的目标是最大化其福利。而供电商收入来源于用户所付电费，成本包含传统能源与新能源供电成本与碳交易成本，则供电商福利可表示如下：

$$\pi^s = E[\sum_{t\in T}\sum_{n\in\mathcal{N}} p_{t,n}\tilde{X}_{t,n} - \sum_{t\in T}(C_t^l(L_t^e) + C_t^{RE}(L_t^r) + C_t^E)] \tag{18}$$

1.3　负荷不确定情况下实时定价模型

考虑社会福利最大化目标，计及负荷不确定情况下的智能电网实时定价模型(19)表述如下：

$$\begin{aligned} &\max\ \mu_1\pi^c + (1-\mu_1)\pi^s \\ &\text{s.t. } (8)\text{–}(9),(13)\ \forall n \in \mathcal{N}, \forall t \in T \end{aligned} \tag{19}$$

其中 $\mu_1 \in (0,1)$， $1-\mu_1$ 分别表示用户侧福利与供电商福利的权重系数。 μ_1 的取值由供电商的定价策略与用户的需求弹性共同决定。可以发现实现最优社会福利时，用户的总负荷与供应商的电力供给是相同的。

1.4　目标函数的转换

目标函数式(19)可以分为用户和供电商两部分，根据期望运算性质，有

$$E[\tilde{X}_{t,n}] = E[X_{t,n} + \delta_{t,n}] = X_{t,n} + E[\delta_{t,n}] = X_{t,n}$$

则目标函数展开表示如下：

$$\begin{aligned} \pi^s &= E[\sum_{t\in T}\sum_{n\in\mathcal{N}} p_{t,n}\tilde{X}_{t,n} - \sum_{t\in T}(C_t^l(L_t^e) + C_t^{RE}(L_t^r) + C_t^E)] \\ &= E[\sum_{t\in T}\sum_{n\in\mathcal{N}} p_{t,n}\tilde{X}_{t,n}] - \sum_{t\in T}(C_t^l(L_t^e) + C_t^{RE}(L_t^r) + C_t^E) \\ &= \sum_{t\in T}\sum_{n\in\mathcal{N}} p_{t,n}X_{t,n} - \sum_{t\in T}(C_t^l(L_t^e) + C_t^{RE}(L_t^r) + C_t^E) \end{aligned}$$

$$\pi^c = E[\sum_{t\in T}\sum_{n\in\mathcal{N}}(U(\tilde{X}_{t,n}) - p_{t,n}\tilde{X}_{t,n})]$$

由前文效用函数定义式(6)可知，效用函数 $U(\tilde{X}_{t,n})$ 的期望：

$$\begin{aligned} E[U(\tilde{X}_{t,n})] &= \beta_n E[X_{t,n} + \delta_{t,n}] - \frac{\alpha_n}{2}E[X_{t,n} + \delta_{t,n}]^2 \\ &= \beta_n X_{t,n} - \frac{\alpha_n}{2}(X_{t,n})^2 - \frac{\alpha_n}{2}\sigma_{\delta_{t,n}}^2 \end{aligned}$$

由上节中所定义随机变量 $\delta_{t,n}$ 的期望与方差定义，令 $\tilde{U}(X_{t,n}) = E[U(\tilde{X}_{t,n})]$，则

$$\tilde{U}(X_{t,n}) = \begin{cases} \beta_n X_{t,n} - \dfrac{\alpha_n}{2}(X_{t,n})^2 - \dfrac{\alpha_n}{2}\sigma_{\delta_{t,n}}^2, & 0 \leqslant X_{t,n} \leqslant \dfrac{\beta_n}{\alpha_n} \\ \dfrac{(\beta_n)^2}{2\alpha_n} - \dfrac{\alpha_n}{2}\sigma_{\delta_{t,n}}^2, & X_{t,n} > \dfrac{\beta_n}{\alpha_n} \end{cases} \tag{20}$$

$$\pi^c = \sum_{t\in T}\sum_{n\in\mathcal{N}}(\tilde{U}(X_{t,n}) - p_{t,n}X_{t,n})$$

于是不确定性模型式(19)可通过期望转换为确定性模型式(21)：

$$\begin{aligned} &\max \sum_{t\in T}\sum_{n\in\mathcal{N}}(\mu_1\tilde{U}(X_{t,n}) + p_{t,n}X_{t,n}) \\ &+(\mu_1 - 1)\sum_{t\in T}(C_t^l(L_t^e) + C_t^{RE}(L_t^r) + C_t^E) \\ &\text{s.t. } (8)\text{–}(9), (13)\ \forall n \in \mathcal{N}, \forall t \in T. \end{aligned} \tag{21}$$

2　算法设计

本节将实时定价模型转换为一种马尔可夫决策过程，基于马尔可夫过程的强化学习能够很好地应用于单智能体环境中，本文使用了一种高效且适应多种环境的Q学习算法进行模型求解。

强化学习(reinforcement learning, RL)是在不同环境中自学习的一种最优动作决策技术[28]，其最重要的特征是智能体学习并记录相应的反馈，目标是最大化智能体的长期累积奖励。智能体通过参数的调整自发选择较大奖励值的动作，具有自我学习与自我更新的优势，交互过程如图 2 所示。

时间差分(temporal-difference learning, TD)算法是强化学习的核心算法，常见的 Q 学习方法就属于 TD 算法，其值函数更新公式为

$$Q(s,a) = Q(s,a) + \partial(r + \gamma Q(s',a') - Q(s,a)) \tag{22}$$

其中 $\partial \in [0,1]$ 是学习率， $\gamma \in [0,1]$ 是折现因子，表明了当前奖励与未来奖励的相对重要性。

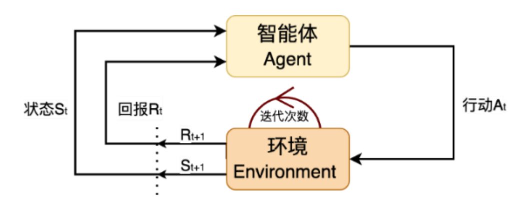


图 2　强化学习中智能体与环境的交互过程

Fig. 2　The interaction process between the agent and environment in reinforcement learning

时间差分算法结合了蒙特卡洛和动态规划(dynamic programming, DP)方法，与蒙特卡洛相似的是可以直接从历史经验中学习。与 DP 类似的是使用后继状态的值函数对当前状态的值函数进行更新。

在每个时间段中，智能体期望最大化累计折扣回报，即最大化当前时段和后续时间段的回报总和，可表述如下：

$$R_t = r_t + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = r_t + \sum_{k=1}^{\infty} \gamma^k r_{t+k+1} \tag{23}$$

强化学习求解最优策略即转换为求状态-动作值函数的最优值。通过实施策略 $A(\lambda_{t,n})$ 将状态 s 转移至状态 s' 而获得转移概率 $P_{ss'}^{A}$ 与回报函数 $R_{ss'}^{A}$，于此本文可以得出最终迭代动作值函数的贝尔曼(Bellman)方程[14]：

$$Q_\pi(s,a) = \sum_{s \in S} P_{ss'}^{a} [R_{ss'}^{a} + \gamma \sum_{a'} Q_\pi(s',a')] \tag{24}$$

其中 $s \in S$ 表示状态集合。

因此，最优策略 A^* 下的最佳状态值函数 $V^*(s)$ 可以表示为

$$V^*(s) = \max_{a \in A} \sum_{s \in S} P_{ss'}^{a} [R_{ss'}^{a} + \gamma \sum_{a'} Q(s',a')] \tag{25}$$

其中 $V^*(s)$ 表示采用最优策略下的状态值函数，a' 表示状态 s' 下所有可能的动作。

在状态转移概率 P 和累积回报 R 已知的情况下，上述 Bellman 最优方程是非线性的，最优策略 $\pi^*(a|s)$ 通常采用迭代的方法求解[29]，根据迭代求解的对象是值函数还是状态动作值函数可以将迭代算法分为值迭代与策略迭代两类。

最终，本文可以得到最优策略为

$$\pi^*(a|s) = \begin{cases} 1, & a = arg\max_{a \in A} Q_\pi(s,a) \\ 0, & a \neq arg\max_{a \in A} Q_\pi(s,a). \end{cases} \tag{26}$$

Q 学习用于求解实时定价模型时，实时电价问题可以表述为马尔可夫决策过程，需要基于马尔可夫决策过程确定强化学习模型要素(S , A , P , γ , R)[30]。通过智能体不断选择针对环境的策略并依据来自环境的反馈逐步迭代，获取到最佳策略，即最优的实时电价是最佳策略的选择过程。供电商根据当前时间段用户侧用电量设置电价即策略，然后用户根据电价从上一状态转移到下一个状态。此转移过程主要取决于当前时段的行动和用户上一个时间段的状态，应用强化学习框架(如图 3 所示)表示供电商与用户之间的能源交易策略，以充分提高社会整体福利。

a)状态空间 S ：定义状态空间时需要综合考虑对决策问题有影响的因素。对于实时定价问题来说，状态空间 S 由负荷需求、负荷和时段组成。 $p_{t,n}$ 代表供电商对用户在 t 时段内提供的电价。 $X_{t,n}$ 表示在用户接收到供电商的价格信号后用户所对应的能源需求量，可视作用户对电价的反馈而实时更新得出的。状态空间集合表示如下：

$$S = \{s \mid s_t = (X_{t,n}, p_{t,n}, L_t^e, L_t^r)\} \tag{27}$$

b)动作空间 A ：由智能体来输出动作即供电商提供的电价 $p_{t,n}$ ，输出的决策动作是一个连续变量，无须离散化操作，因此，本节将动作空间设置为一个连续的电价区间范围。

$$A = \{a \mid a \in [c_n^{\min}\pi_0, c_n^{\max}\pi_0]\} \tag{28}$$

c)状态转移概率 P ：对应式(24)，定义实时定价策略下状态转移概率 $P \in P_{s_t s_{t+1}}^{a}$ ， $P_{s_t s_{t+1}}^{a}$ 表示为智能体在状态 s_t 下采取动作 a 后将会环境转移到下个阶段 s_{t+1} 的转移概率。

d)折现因子 γ ： γ 是折现因子，指当前决策动作下未来奖励期望所占的比例。一般来说， γ 越大，未来奖励相较于当前奖励的重要程度越高，当前时段的决策将对下一状态产生重要的影响，若折现率为 0 即只考虑当前奖励将会造成算法的“短视”优化。

e)回报 R ：在本节中，实时定价模型考虑社会福利最大化作为目标，将回报与社会福利值对应，因此单一阶段的具体回报定义如下：

$$\begin{aligned} r_t = & \sum_{n \in \mathcal{N}} (\mu_1 \tilde{U}(X_{t,n}) + p_{t,n} X_{t,n}) \\ & + (\mu_1 - 1) \sum_{t \in T} (C_t^l(L_t^e) + C_t^{RE}(L_t^r) + C_t^E) \end{aligned} \tag{29}$$

综上，实时定价策略下的 Q 值函数更新式如下：

$$Q^k(s_t,a_t) \leftarrow (1-\partial) Q^{k-1}(s_t,a_t) + \partial (r_t + \gamma Q^{k-1}(s_{t+1},a_{t+1})) \tag{30}$$

其中 $\partial \in [0,1]$ 是学习率。

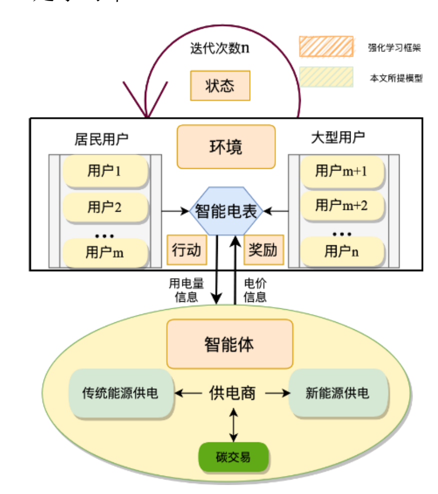


图 3　强化学习实时定价框架

Fig. 3　Real-time pricing mechanism based on reinforcement learning

在迭代开始即 $t=0$ 时，模型的目标是最大化当天所有时段的总效益。第一个时段结束后，目标将转换为最大化剩余时段的总奖励。在每个时间段的末尾最大化一天中剩余时段的奖励，可充分考虑到时间的前后关联性，Q 学习实时定价机制如下。

算法 1　Q 学习实时定价机制

输入：预设参数，初始负荷值 $X_{t,n}^{0}$ 、供电量 $L_t^{e,0}, L_t^{r,0}$ 和电价 $p_{t,n}^{0}$ 。

输出：最优动作值函数 $Q^{\pi *}$ ，负荷 $X_{t,n}^{*}$ ，最优供电量 $L_t^{e,*}, L_t^{r,*}$ 与最优电力价格 $p_{t,n}^{*}$ 。

a）数据初始化，初始化动作值函数 $Q^0(s,a)=0, k=0, t=0$ ；

b）迭代： $k \leftarrow k+1$ ；

(a)对每一轮循环，重复 $t \leftarrow t+1$；

(b)如果 $|Q^k - Q^{k-1}| \leqslant \delta$ 成立，停止迭代输出 Q^k 。否则转(c)；

(c)面对初始策略，观察状态 s_t 并选择一个动作 a_t ；

(d)智能体观察收益值函数 r_t ，以及观察下一个状态 s_{t+1} ；

(e)更新动作值函数

$Q^k(s_t,a_t) \leftarrow (1-\partial) Q^{k-1}(s_t,a_t) + \partial (r_t + \gamma Q^{k-1}(s_{t+1},a_{t+1}))$ ；

(f)检查是否完成一个周期，如果 $t=T$ ，跳出循环。否则转(g)；

(g)通过式(2)(7)和(28)计算出实时电价、供电量与负荷。

强化学习寻优常见的方法是使用 $\varepsilon - greddy$ 策略[31]，此策略可选择具有给定概率分布的随机动作。在一天开始时，智能体即供电商首先在给定状态的价格边界内随机选择初始策略 a_0 即初始供电价格。选择初始策略后，智能体可以立即获得一定的奖励，同时智能体还将观察时段中环境并更新

Q 值即社会福利值。随着学习深入与供电商反复的价格调整，Q 值通过智能体与环境学习而增加最终收敛到最大值。当 Q 学习算法实现了足够多的状态与动作后，算法可以保证模型收敛至最优函数[32]。当 $|Q^k - Q^{k-1}| \leqslant \delta$ 时，满足终止条件，模型将收敛至最优值即最大社会福利值，同时获得最优的状态空间。

3　数值仿真

3.1　算例背景

本节介绍数值仿真实验，以验证模型的合理性与算法的有效性。假设某个区域存在供电商与一个社区，考虑了含传统能源与新能源发电的供电商与包含 20 个居民用户与 5 个大型用户的社区，而智能电表可以通过聚合同类型用户的用电信息进行统一调度从而有效保护用户的隐私。本文考虑基于典型日的光伏和风电出力，如附录中图 A1 所示。因此，直接参与电力交易的一天内是不同类型用户的总负荷。本文采用文献[33]中的居民及大型用户负荷数据并按照相应的比例进行调整作为本文数据来源，两类用户各个时段的负荷需求见图 A2 和图 A3。

实验环境设置如下：Intel 8259U，RAM 8G， Windows 10 操作系统，Python 3.9 作为编程环境。算例的详细参数见附录中表 3~6，价格弹性系数见附表 3，碳交易价格即碳交易市场中单位碳排放权的价格 p_c，取基准方案下每吨 130 元[27]。考虑到不同用户对于电价的不同反应，对不同类型用户设置不同的效用参数[32]，用户效用参数 β_n 服从均匀分布，用户侧模型参数设置详见表 4。强化学习算法初始参数值设置及供电侧各类参数见附表 5，权重系数 μ_1 由算法自适应选取。同时，本文同时考虑将上海市分时电价与所提实时定价模型进行对比，分时电价见表 6。

3.2　结果分析

用户侧实时电价与负荷削减量分别如图 4 和 5 所示，从图 4 中也可以看出两类用户实时电价趋势相同。将高峰时段(如 10:00-15:00，18:00-21:00)与非高峰时段(如 21:00-7:00)相比较，可以发现高峰时段的用户电价变化率与负荷削减比率高于非高峰时段，这是由于高峰时段电力价格弹性系数较高，价格的变化对于需求侧削峰填谷具有更好的效果，供电侧可在较小的电力价格调整下取得较大的调控力度，同时价格区间约束使电价在保持合理的范围。图 5 表示两类用户的负荷总削减量，从图 5 可以发现大型用户的负荷削减量大于居民用户，这是由于价格区间约束大型用户具有较高的电价且在高峰期的电价波动性较高。

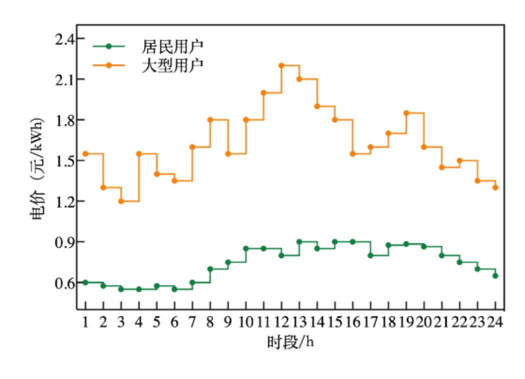


图 4　用户侧实时电价

Fig. 4　Real-time electricity prices for the user side

图 6 为用户侧福利值，可以看出大型用户福利值高于居民用户福利值，同时大型用户电价在用电高峰期间变化率较大，即用户参与负荷调控的意愿较高。用户面对供电商电价的变化按照福利最大化目标调整自身负荷。

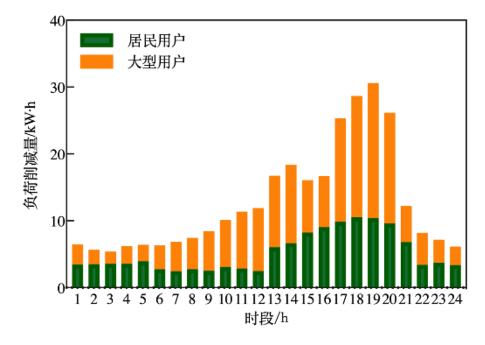


图 5　用户侧负荷削减量

Fig. 5　Load reduction of the user side

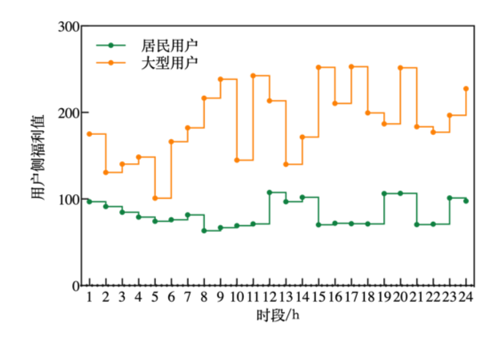


图 6　用户侧福利值

Fig. 6　Welfare values of the user side

图 7 与图 8 反映了供电商最终的供电量、供电商福利与碳交易成本，当取得最优社会福利时用户总负荷与供电商总供电量相同。在考虑碳交易的情况下，供电商优先使用风电、光伏等新能源供电，在缓解化石能源供电压力的同时降低了发电成本，图 8 中碳排放成本为负值，即碳交易能够增加供电商福利，供电商通过新能源发电获得的碳排放权配额超出实际总碳排量，有效提高了供电侧福利。算例验证了碳交易下模型的合理性与有效性，同时碳交易的普及能够有效推进能源系统绿色发展，从而在社会层面促进新能源的有效消纳。

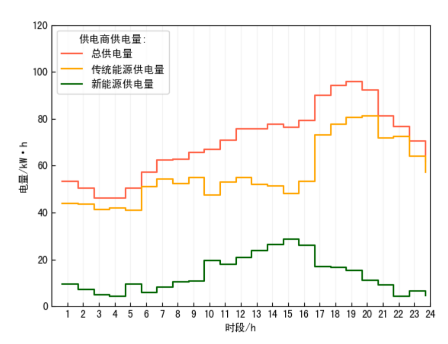


图 7　总供电量，传统能源与新能源供电商供电量

Fig. 7　Total power supply, the amount of power supplied by traditional and new energy suppliers

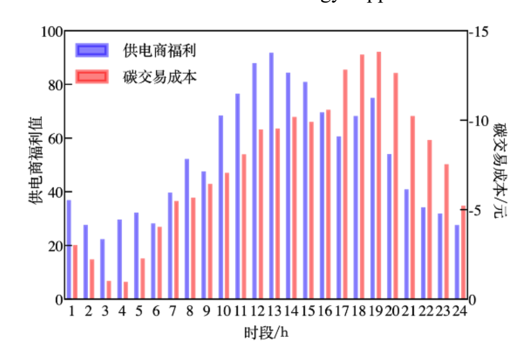


图 8　供电侧福利值与碳交易成本

Fig. 8　Welfare values and carbon trading costs of the supply side

为进一步对比所提模型的合理性与有效性。将本文所提实时定价场景(场景 1)与下面三种场景进行比较，第一种场景为考虑负荷确定情况下的实时定价策略(场景 2)、第二种场景在第一种场景基础上考虑算法的“短视”优化情景(场景 3)、第三种场景为基于分时定价机制下的社会福利(场景 4)，共四种场景。

表 2 显示了一个典型日内四种不同情景下的模型指标值。为了进一步说明所提出考虑负荷不确定下实时定价模型的合理性，假设四种场景基础参数相同。仿真得出所提实时定价场景与另外三种情形的社会福利值类似，同时在实时定价下的社会福利值总是优于分时电价下的社会福利值。尽管不确定情况下的实时定价相较于确定性情况下福利值有所降低，但不确定情况下的实时定价能够更加贴合用户实际用电情况。即所提实时定价策略在保证模型鲁棒性的情况下取得较优的社会福利值，场景对比验证了所提模型的有效性与合理性。

表 2 四种场景下模型指标值

Tab. 2 Model indicator values in four types of scenarios

场景	社会福利值	用户侧福利值	供电商福利值	碳交易成本/元
场景 1	2291.5	5048.9	1372.4	-262.5
场景 2	2311.9	5106.1	1380.5	-265.5
场景 3	2016.5	4679.5	1128.8	-150.3
场景 4	2146.7	4795.8	1263.7	-113.8

4 结束语

本文使用强化学习框架 Q 学习算法求解实时电价，算例仿真验证了所提策略的有效性，并具有下列优势：a)本文应用强化学习框架将实时定价问题转换为一个马尔可夫决策过程，供电商作为智能体可在与全体用户迭代过程中学习和获取最优的实时定价策略，实现电价的自动优化。b)本文考虑用户分类，可有效地提升系统性能同时符合用户实际用电情况。c)Q 学习算法可适用于所提实时电价模型求解。计及负荷不确定的实时定价策略能够有效地平衡电力市场能源供需，提高电力系统的鲁棒性。d)碳排放交易机制能够有效助力“双碳”目标的实现，使得供电侧在优化调度中充分调用风电、光伏等可再生能源，充分提高了电力系统的经济性与环保性。

本文提出的策略可以使用多种方式扩展。后续可以引入用电限制与用户资金限制等约束条件,从而更加贴近现实情况；运用多智能体强化学习算法整合含电动汽车与储能设备的区域能源微网方案，从而提高强化学习框架对于复杂电力系统的适应性；针对更大规模用户，通过大数据驱动分布式强化学习可实现更优的电力需求侧管理。

5 附录

表 3 价格弹性系数

Tab. 3 Elasticity of demand

用户类型	非高峰期 (21:00-7:00)	中峰 (7:00-10:00) (15:00-18:00)	高峰 (10:00-15:00) (18:00-21:00)
居民用户	0.3	0.5	0.8
大型用户	0.1	0.15	0.25

表 4 用户侧参数设置

Tab. 4 User-side parameter setups

参数	居民用户	大型用户
$(c_n^{\min}, c_n^{\max})$	(1,2)	(2.5,5)
β_n	[3,4]	[5,8]
α_n	0.14	0.02
$\sigma_{t,n}$	5	10

表 5 供电侧及强化学习参数设置

Tab. 5 Power supply side and RL parameter setups

参数	值	参数	值
$a_t / b_t / c_t$	0.01/0/0	v_{rated}	20
$L_t^{e,\min}$	40	v_{in} / v_{out}	3/25
$L_t^{e,\max}$	200	δ_e / δ_r	0.7/1
δ_{RE}	0.005	σ_{RE}	0
π_0	0.45	p_e	130
P_{PV}^{rated}	400	α_e	0.0034
P_{WT}^{rated}	400	β_e	-0.38
G_{PV}	65	λ_e	36
η_{PV}	0.093	N_{PV}	1
$T_{PV,T}$	25	N_{WT}	1
∂ / δ	0.9/0.05	γ	0.9

表 6 分时电价设置

Tab. 6 TOU pricing setups

用户类型	分时电价	
	峰时(6:00-22:00)	谷时(22:00-6:00)
居民用户	0.617	0.307
大型用户	1.145	0.562

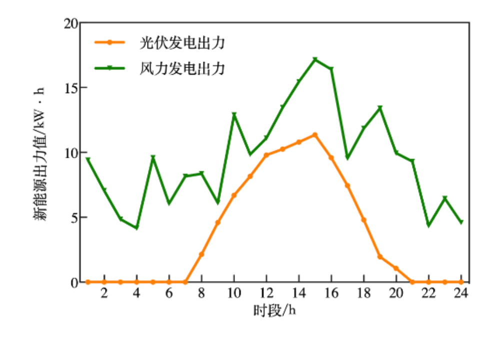


图 9 典型日的新能源供电商风光出力值

Fig. 9 Renewable power supplier’s outputs of typical day

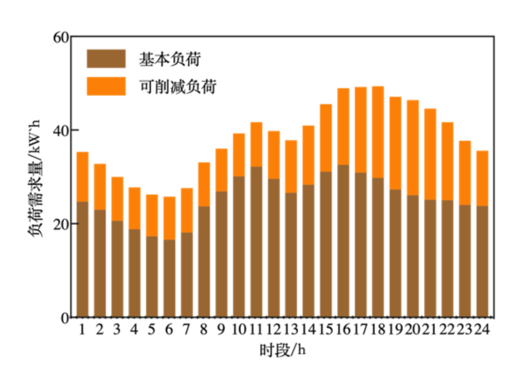


图 10 居民用户负荷需求

Fig. 10 Load demand of residential users

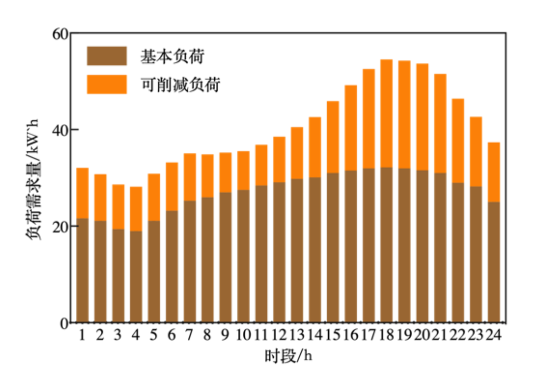


图 11 大型用户负荷需求图

Fig. 11 Load demand of large use

参考文献：

[1] 张瑶，王傲寒，张宏. 中国智能电网发展综述 [J]. 电力系统保护与控制, 2021, 49 (5): 180-187. (Zhang Yao, Wang Aohan, Zhang Hong. Overview of smart grid development in China [J]. Power System Protection and Control, 2021, 49 (5): 180-187.)

[2] 黄开艺，艾芊，张宇帆，等. 基于能源细胞-组织架构的区域能源网需求响应研究挑战与展望 [J]. 电网技术, 2019, 43 (9): 3149-3160. (Huang Kaiyi, AI Qian, Zhang Yufan, *et al.* Challenges and prospects of regional energy network demand response based on energy cell-tissue architecture [J]. Power System Technology, 2019, 43 (9): 3149-3160.)

[3] 原冠秀，高岩，王宏杰. 基于效用分类的智能电网实时电价算法 [J]. 上海理工大学学报, 2020, 42 (1): 29-35. (Yuan Guanxiu, Gao Yan, Wang Hongjie. A real-time pricing algorithm based on utility classification in a smart grid [J]. Journal of University of Shanghai for Science and Technology, 2020, 42 (1): 29-35.)

[4] Samadi P, Mohsenian-Rad A H, Schober R, *et al.* Optimal real-time pricing algorithm based on utility maximization for smart grid [C]// IEEE International Conference on Smart Grid Communications, Piscataway, NJ: IEEE Press, 2010: 415-420.

[5] 王宏杰，高岩. 基于非光滑方程组的智能电网实时定价 [J]. 系统工程学报, 2018, 33 (03): 320-327. (Wang Hongjie, Gao Yan. Research on the real-time pricing of smart grid based on nonsmooth equations [J]. Journal of Systems Engineering, 2018, 33 (03): 320-327.)

[6] 陶莉，高岩，朱红波. 以极小化峰谷差为目标的智能电网实时定价 [J]. 系统工程学报, 2020, 35 (03): 315-324. (Tao Li, Gao Yan, Zhu Hongbo. Real-time pricing strategy for smart grid based on the minimization of the peak-valley difference [J]. Journal of Systems Engineering, 2020, 35 (03): 315-324.)

[7] 李军祥，周继儒，何建佳. 基于区块链的电网实时定价混合博弈研究 [J]. 电网技术, 2020, 44 (11): 4183-4191. (Li Junxiang, Zhou Jiru, He Jianjia. Mixed game of real-time pricing based on block chain for power grid [J]. Power System Technology, 2020, 44 (11): 4183-4191.)

[8] Yuan Guanxiu, Gao Yan, Ye Bei, *et al.* Real-time pricing for smart grid with multi-energy microgrids and uncertain loads: a bilevel programming method [J]. International Journal of Electrical Power & Energy Systems, 2020, 2020 (123): 106206.

[9] 高岩. 智能电网实时电价社会福利最大化模型的研究 [J]. 中国管理科学, 2020, 28 (10): 201-209. (Gao Yan. The social welfare maximization model of real-time pricing for smart grid [J]. Chinese Journal of Management Science, 2020, 28 (10): 201-209.)

[10] Tao Li, Gao Yan. Real-time pricing for smart grid with distributed energy and storage: a noncooperative game method considering spatially and temporally coupled constraints [J]. International Journal of Electrical Power & Energy Systems, 2020, 2020 (115): 105487.

[11] 朱红波，高岩，后勇，等. 马尔可夫过程下多类用户智能电网实时电价 [J]. 系统工程理论与实践, 2018, 38 (3): 807-816. (Zhu Hongbo, Gao Yan, Hou Yong, *et al.* Real-time pricing considering different type of users based on Markov decision processes in smart grid [J]. Systems Engineering-Theory & Practice, 2018, 38 (3): 807-816.)

[12] Mnih V, Kavukcuoglu K, Silver D, *et al.* Human-level control through deep reinforcement learning [J]. Nature, 2015, 518 (7540): 529-533.

[13] José R, Zoltán N. Reinforcement learning for demand response: A review of algorithms and modeling techniques [J]. Applied Energy, 2019, 2019 (235): 1072-1089.

[14] Lu Renzhi, Hong SeungHo, Zhang Xiongfeng. A dynamic pricing demand response algorithm for smart grid: reinforcement learning approach [J]. Applied Energy, 2018, 2018 (220): 220-230.

[15] Zhang Li, Gao Yan, Zhu Hongbo, *et al.* Bi-level stochastic real-time pricing model in multi-energy generation system: A reinforcement learning approach [J]. Energy, 2021, 2021 (239): 121926.

[16] 冯小峰，谢添阔，高赐威，等. 电力现货市场下计及售电商长期收益的需求侧响应 [J]. 电网技术，2019, 43 (08): 2761-2769. (FENG Xiaofeng, XIE Tiankuo, GAO Ciwei, *et al.* A demand side response strategy considering long-term revenue of electricity retailer in electricity spot market [J]. Power System Technology, 2019, 43 (8): 2761-2769.)

[17] Wang Huiwei, Huang Tingwen, Liao Xiaofeng, *et al.* Reinforcement learning in energy trading game among smart microgrids [J]. IEEE Trans on Industrial Electronics, 2016, 63 (8): 5109-5119.

[18] Du Yan, Li Fangxing, Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning [J] , IEEE Trans on Smart Grid, 2020, 11 (2): 1066–1076

[19] Jin Ming, Feng Wei, Marnay C, *et al.* Microgrid to enable optimal distributed energy retail and end-user demand response [J]. Applied Energy, 2018, 2018 (210): 1321-1335.

[20] 张莉，高岩，朱红波，等. 考虑用电量不确定性的智能电网实时定价策略 [J]. 电网技术, 2019, 431 (10): 181-190. (Zhang Li, Gao Yan, Zhu Hongbo, *et al.* Real-time pricing strategy based on uncertainty of power consumption in smart grid [J]. Power System Technology, 2019, 431 (10): 181-190.)

[21] Tarasak P. Optimal real-time pricing under load uncertainty based on utility maximization for smart grid [C]// IEEE International Conference on Smart Grid Communications, Piscataway, NJ: IEEE Press, 2011: 321-326.

[22] Yu Mengmeng, Hong Seung Ho. Incentive-based demand response considering hierarchical electricity market: A Stackelberg game approach [J]. Applied Energy, 2017, 2017 (203): 267-279.

[23] Samadi P, Mohsenian-Rad H, Schober R, *et al.* Advanced demand side management for the future smart grid using mechanism design [J]. IEEE Trans on Smart Grid, 2012, 3 (3): 1170-1180.

[24] 李军祥，潘婷婷，高岩. 智能电网互补能源供用电实时定价算法研究 [J]. 计算机应用研究, 2020, 37 (4): 1092-1096. (Li Junxiang, Pan Tingting, Gao Yan, Real time pricing algorithm for supply and demand of complementary energy on smart grid [J]. Application Research of Computers, 2020, 37 (4): 1092-1096.)

[25] Chiu Techuan, Shih Yuanyao, Pang Aichun, *et al.* Optimized day-ahead pricing with renewable energy demand-side management for smart grids [J]. IEEE Internet of Things Journal. 2017, 4 (2): 374-383

[26] Zhang Ning, Hu Zhaoguang, Dai Daihong, *et al.* Unit commitment model in smart grid environment considering carbon emissions trading [J]. IEEE Trans on Smart Grid, 2016, 7 (1): 420-427.

[27] 张晓辉，梁军雪，赵翠妹，等. 基于碳交易的含燃气机组的低碳电源规划 [J]. 太阳能学报, 2020, 41 (07): 92-98. (Zhang Xiaohui, Liang Junxue, *et al.* Research on low-carbon power planning with gas turbine units based on carbon transactions [J]. Acta Energiae Solar Sinica, 2020, 41 (07): 92-98.)

[28] Alpaydin E. Introduction to machine learning [M]. 4th ed. Cambridge: MIT press, 2020.

[29] Yu Tao, Zhou Bin, Chan Kawing, *et al.* Stochastic optimal relaxed automatic generation control in non-markov environment based on multi-step Q (λ) learning [J]. IEEE Trans on Power Systems, 2011, 26 (3): 1272-1282.

[30] Kong Xiangyu, Kong Deqian, *et al.* Online pricing of demand response based on long short-term memory and reinforcement learning [J]. Applied Energy, 2020, 2020 (271): 114945.

[31] Han Xuefeng, He Hongwen, Wu Jingda, *et al.* Energy management based on reinforcement learning with double deep Q-learning for a hybrid electric tracked vehicle [J]. Applied Energy, 2019, 2019 (254): 113708.

[32] Hasselt H. Double Q-learning [J]. Advances in neural information processing systems, 2010, 2010 (23): 2613-2621.

[33] Yang Peng, Tang Gongguo, Nehorai A, A game-theoretic approach for optimal time-of-use electricity pricing [J]. IEEE Trans on Power Systems, 2012, 28 (2): 884-892.

[34] Lin Jie, Xiao Biao, Zhang Hanlin, *et al.* A novel multitype-users welfare equilibrium based real-time pricing in smart grid [J]. Future Generation Computer Systems, 2020, 2020 (108): 145-160.